DOCUMENT RESUME

ED 318 758 TM 014 794

AUTHOR Reckase, Mark D.
TITLE Unidimensional Data from Multidimensional Tests and Multidimensional Data from Unidimensional Tests.
SPONS AGENCY Office of Naval Research, Arlington, Va.
PUB DATE Apr 90
CONTRACT N00014-89-J-1908
NOTE 47p.; Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Item Response Theory; Matrices; *Statistical Analysis; Test Construction; *Test Results
IDENTIFIERS *Multidimensional Approach; Psychological Dimensionality; Statistical Dimensionality; *Unidimensionality (Tests)

ABSTRACT

Although the issue of dimensionality of the data obtained from educational and psychological tests has received considerable attention, the terms "unidimensional" and "multidimensional" have not been used very precisely. One use of the term dimensionality is to refer to the number of hypothesized psychological constructs believed to be required for successful performance on a test; this is "psychological dimensionality." A second use is to refer to the minimum number of mathematical variables needed to summarize a matrix of item response data; this is "statistical dimensionality" (SD). This paper presents a framework for discussing the SD of dichotomous data matrices of item response data. It is suggested that dimensionality refers to the number of dimensions needed to summarize a data matrix that results from interactions of a set of test items and a group of examinees. Two cases are presented to summarize the circumstances under which a set of test items will generate unidimensional data when responded to by a population of individuals. In the first case, all items are sensitive to the same combination of skills in the same manner. The second situation is when psychological dimensions are strongly confounded with the difficulty of the test items. Fourteen perspective, contour, or vector plots illustrate the discussion. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

MARK D. RECKASE

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED318758

Unidimensional Data from Multidimensional Tests and Multidimensional Data from Unidimensional Tests

Mark D. Reckase
American College Testing (ACT)

Paper presented at the annual meeting of the American Educational Research Association, Boston, April, 1990. This research was supported by Contract No. N00014-89-J-1908 with the Cognitive Science Research Programs of the Office of Naval Research.

TM014794

Unidimensional Data from Multidimensional Tests and Multidimensional Data from Unidimensional Tests

In the current literature on testing, the issue of the dimensionality of the data obtained from educational and psychological tests has received considerable attention (e.g., Goldstein & Wood, 1989; Hulin, Drasgow & Parsons, 1983; Linn, 1989). This interest has probably been stimulated by the development of item response theory (IRT) since most IRT models assume that the construct underlying test performance is unidimensional (Hambleton, 1989). However, the terms "unidimensional" and "multidimensional" have often not been used very precisely, resulting in some confusion. The purpose of this paper is to make a proposal for how these terms should be used and to provide some examples showing what dimensionality really means when applied to dichotomous item response data.

There are two frequent uses of the term dimensionality when the term is used in reference to psychological and educational tests. First, dimensionality is used to refer to the number of hypothesized psychological constructs that are believed to be required for successful performance on a

test (Embretson, 1985). For example, numerical computation and verbal reasoning are said to be required to successfully perform on a mathematics story problem. Numerical computation and verbal reasoning are two psychological dimensions that are hypothesized to exist to explain differences in performance on the test item. In this paper, this use of dimensionality will be referred to as "psychological dimensionality," or PD for short. Second, dimensionality is used to refer to the minimum number of mathematical variables that is needed to summarize a matrix of item response data. For example, a vector composed of two elements may be needed in a probabilistic model of test performance to reasonably accurately predict how a person will respond to a particular set of test items. In this paper, this use of the term will be referred to as "statistical dimensionality," or SD for short.

The meaning of the constructs being used in these two cases may or may not be the same. That is, differences in level on the mathematical variables may not translate directly into differences on the psychological

constructs. Whether or not they have the same meaning is a question of the validity of the measures obtained using the particular mathematical model of the interactions of persons and test items.

An issue related to the use of the term "dimensionality" is to what, precisely, does the term refer? Does a psychological or educational test have a particular dimensionality? Or does the dimensionality reside in the examinee population? Can the dimensionality of a test and an examinee population be different? In psychological and educational testing, the data typically analyzed is a matrix of zeros and ones that is generated by of the interaction of a set of persons and a set of test items. It is that matrix of observed data that is analyzed to determine the level of dimensionality, specifically the SD. Therefore, the SD is a characteristic of the data matrix, not the test or the examinee population. It cannot be said on the basis of statistical analyses that a test is statistically unidimensional or multidimensional, for if the test were administered to a different population of examinees, a different SD might result. The observed SD is a function

of the test and examinee population, but it is not necessarily descriptive of a characteristic of either the test or the examinees.

Given this discussion of the use of the term dimensionality, the purpose of this paper is to present a framework for discussing the SD of dichotomous data matrices composed of item response data and to present some theory to help interpret the meaning of such phrases as "these test data can be adequately modeled with a unidimensional model" or "a multidimensional model is needed to describe the interaction between the examinee population and this set of test items."

Some Theory for Discussing the Dimensionality of Test Data

Much of psychometric theory has been devoted to determining means for converting the item response matrix of zeros and ones into one or more numerical values that provide meaningful information about the skills possessed by individuals who responded to the items. In the most general sense, this paper is about the circumstances under which it is reasonable to

report a single numerical value to summarize level of performance on a test and when more than one value is needed to provide the summary. The probabilistic modeling procedure called multidimensional item response theory (MIRT) (Reckase, 1989) will be used as the framework for these discussions, but in many cases the results are much more general than the particular models used. Assumptions are made at the start that all matrices generated by the interaction between people and test items are based on psychological processes that have more than a single component and that the probability of a correct response to an item increases with increases in relevant psychological dimensions. This does not necessarily imply that the SD of a data matrix is more than one. To support this contention, it is now necessary to get into the theory.

Let us first begin with a very simple case. Suppose that a population of people vary on only two skills labeled θ_1 and θ_2. On all other skills that they possess, there are no differences. In this case, θ_1 and θ_2 are considered as psychological constructs and not necessarily statistical

constructs. Suppose further that one, dichotomously-scored, test item that requires both of these skills for successful performance is administered to the population. To discuss the relationship between the population and the test item in any quantitative sense requires that a metric be defined for the θ_1,θ_2-space. One way to do this is to specify a mathematical model for the relationship between performance on the item and the location in the space. The model specification sets the metric of the space. The location of the θ_1,θ_2-points relative to each other are determined in such a way that the observed data conform to the relationship specified in the model.

When dealing with only one test item, few constraints are placed on the choice of a model. As more test items are included in the generation of the data matrix, many models will not yield metrics for the space that are consistent with these data. The symptom of this problem is lack of fit of the model to the data matrix.

While many MIRT models are possible (see Reckase and McKinley, 1985 for some examples), the single item in this example will be modeled

using the multidimensional extension of the two-parameter logistic model because of its mathematical tractability and simplicity. This model is given by the equation

$$P\{u_{ij}=1|a_i, d_i, \theta_j\}=\frac{e^{(a_i'\theta_j+d_i)}}{1+e^{(a_i'\theta_j+d_i)}}$$

where $P\{u_{ij}=1|\underline{a}_i, \underline{d}_i, \theta_j\}$ is the probability of a correct response for person $\underline{j}$ to item $\underline{i}$,

$\underline{u}_{ij}$ is the score (0 or 1) for person $\underline{j}$ on item $\underline{i}$,

$\mathbf{a}_i$ is a vector of discrimination parameters,

θ_j is a vector of person location parameters,

and $\underline{d}_i$ is a scalar parameter related to the difficulty of the item.

Alternatively, the metric of the θ_1,θ_2-space could be defined on some theoretical grounds (e.g., according to theory, the distribution of skills should be standard bivariate normal with $\rho = .7$). In that case, the form of the relationship between the two hypothetical constructs and performance

on the test item would have to be determined empirically from the observed responses to the test item while constraining the bivariate ability distribution to the form that was assumed.

In the model that is used here, the height above the θ_1,θ_2-plane is an estimate of the probability that a person at a particular location in the space will get the test item correct. Another interpretation that could be used is that the height represents the proportion of persons at that point that would obtain a correct response. This interpretation is less defensible since if the space is continuous the number of persons at a point in the space is zero.

If the item under discussion is given some concrete characteristics, the surface describing the probability of a correct response to the item, called the item response surface (IRS), can be presented graphically. Suppose the test item has parameters $\mathbf{a}_1$=(1.732, 1.000) and $\underline{d}_1$=2.8. For convenience, this test item will be labeled Item 1. The surface representing the relationship specified by Equation 1 can be presented in two forms. One is as a three-dimensional perspective representation as is

shown in Figure 1. The second is as a two-dimensional plot of the equiprobable contours as is shown in Figure 2.

Insert Figures 1 and 2 about here

Notice that in Figure 2 the equiprobable contour lines are all parallel, straight lines. This feature is a result of the compensatory nature of the model presented in Equation 1. The $a_i'\theta_j$ term in the exponent of e can be written as $a_{i1}\theta_{j1} + a_{i2}\theta_{j2}$. In this form it is clear that any change in θ_{j1} can be compensated for by a change in θ_{j2} resulting in the same value for the sum. In fact, the equation

$$a_{i1}\theta_{j1} + a_{i2}\theta_{j2} + d_i = k$$

is the equation for the linear contour lines with the value of k determining which line is being considered. Note that the slope of each of the equiprobable contours for Item 1 with respect to θ_1 is -1.732.

The Dimensionality of Data from a Single Test Item

By definition, the probability of a correct response to the test item for persons located along the same equiprobable contour line is the same. This implies that all persons along the same contour line can be mapped into the same numerical score without loss of information about their probability of successfully completing the test item. For this reason, the responses to all dichotomously-scored, one-item tests have a SD of one. This result is not a function of the compensatory MIRT model used to model the IRS; the result is much more general. Even if the contour lines are curved, or even discontinuous, a mapping still exists from the θ-plane to the 0-1 region of the number line. The result is simply a function of the dichotomous scoring of the test item.

This does not mean that the number of PDs needed to respond to the item is one. All test items probably require more than one PD to determine the correct response. The richness of skills required to correctly respond to the test item is lost in the conversion of performance on the test item into a zero or a one. That is the price paid for convenience of

0/1-scoring.

The fact that the interaction of a population and a single dichotomously-scored, test item generates data that has a SD of one does not trivialize the question of what the one dimension is in terms of the PDs. Note that in the contour plot shown in Figure 2, the probability of a correct response changes more quickly as θ_1 increases than it does when θ_2 increases. Thus, the test item provides better discrimination between people who differ on the first psychological dimension than on the second. Another way of showing this is with a vector from the origin of the space with its base on the .5-contour, pointing in the direction of greatest rate of change in probability for the IRS. Such a vector for Item 1 is shown in Figure 3. The length of the vector is related to the steepness of the slope of the probability surface in the direction specified. Note that the vector is pointing more along θ_1 than θ_2, indicating that Item 1 is more sensitive to differences in θ_1 than differences in θ_2. The direction of the vector tells the

relative weighing of the two psychological dimensions on performance on the test item.

Insert Figure 3 about here

The Dimensionality of Data from a Two-item Test

A test that is composed of two test items provides a much more interesting situation for study. Three different cases will be considered.

Case 1

Suppose that the test is composed of the test item already discussed and a second test item (Item 2) that has equiprobable contours that are everywhere parallel to those for Item 1. The parameters of one such item are $\underline{a}_2$=(1.299, .75) and d_2=-1.35. The response surface plots for Item 2 are given in Figures 4 and 5 for the perspective plot and the contour plot, respectively.

Insert Figures 4 and 5 about here

By comparing these plots with those in Figures 1 and 2, it can easily be seen that the contour lines are parallel. This can be shown analytically since the equation for a contour line for Item 2 is $1.299\theta_1 + .75\theta_2 - 1.35 = \underline{k}$. The slope of any of these contour lines with respect to θ_1 is $-1.299/.75 = -1.732$, the same slope as for the contours of Item 1. However, the magnitude of the probability of a correct response for a particular person is not the same for the two items. All persons having the same probability of a correct response to Item 1 may have a different probability of correct response to Item 2, but if two persons had equal probabilities of correct response to Item 1, the two persons will also have equal probabilities of correct response to Item 2.

An expected score on the two item test can be obtained for each person in the population by summing the probability of correct response to

the two test items. This process is based on the assumption that local independence holds. The contour surface in Figure 6 shows the equal expected score contours for this two item test. Note that these contours are also parallel to equiprobable contours for the two test items. This obviously must be the case because the expected score for person j, $\underline{x}_j$, is simply the sum of the probabilities of correct response for the two items

$$E(x_j|\theta_j) = P(u_{1j}|a_1, d_1, \theta_j) + P(u_{2j}|a_2, d_2, \theta_j).$$

Since both the values of $P(u_{1j}|a_1, d_1, \theta_j)$ and $P(u_{2j}|a_2, d_2, \theta_j)$ are constant along a contour, the value of $E(x_j|\theta_j)$ must also be constant along the contour.

Insert Figure 6 about here

The fact that all of the persons on an equiprobable contour for one item are on an equiprobable contour for the other item, and have the same

expected score on the two item test, indicates that a simpler model than the two dimensional MIRT model can be used to describe the item-person interactions in this special case. One such model is to project each θ_1,θ_2-point onto a line orthogonal to the equiprobable contours. A single number, the distance along this line from an arbitrary origin, can be used to obtain the probability of correct response to each item and the estimated number-correct score on the two-item test. Actually, any line that intersects the equiprobable contours could be used, but the orthogonal line is more convenient mathematically. Since the performance of all of the persons on the same equiprobable contour can be modeled using a single numerical value without loss of precision, the data generated by the interaction with these two test items has a SD of one. Of course, that fact does not mean that only one psychological trait was required to successfully respond to the test items. In this case the example was developed based on the premise that two psychological dimensions were required.

Figure 7 shows the vectors for both Item 1 and 2. Note that they fall along the same straight line. This indicates that both test items are sensitive to differences in the two psychological dimensions in the same way. Because the vectors provide a more concise summary of the probability surfaces, they are often more useful for describing the characteristics of a set of test items than the perspective or contour plots.

Insert Figure 7 about here

Case 2

Suppose that the second test item does not have contour lines that are parallel to those from Item 1. This situation is shown in Figures 8 by the contour plot for a third test item. The test item shown in this figure, henceforth called Item 3, has item parameters $\mathbf{a}_3$=(.166, 1.893) and $\underline{d}_3$=.789. Note that persons on the same contour line for Item 1 will have quite different probabilities of a correct response for Item 3. The expected

scores for the two-item test constructed from Items 1 and 3 are shown by the contour plot in Figure 9. Figure 10 shows the vector plot for the two item test. Unlike those in Figure 7, the vectors in this plot clearly do not fall on the same line. Note that the contours in Figure 9 are not parallel to either those of Item 1 or Item 3. Clearly the data generated by the interaction of the population and these two test items will have an SD greater than one. The probability of correct response to Item 3 for persons along an equiprobable contour for Item 1 vary substantially. Mapping all points on an equiprobable contour into a point on a line will not yield the simplification that was obtained in Case 1.

Insert Figures 8, 9, and 10 about here

Case 3

The two cases that have been discussed up to this point have varied two characteristics of the items in the two-item tests. In Case 1, the items varied in difficulty, but the direction of best measurement in the space was the same (the vectors pointed in the same direction). In Case 2, the two items were of about the same difficulty, but performance on the different items required quite different weightings of the skills in the psychological space. A third possibility is that the items vary both on their difficulty and on the weighing of the dimensions needed to solve the test items. That case, Case 3, is shown by the two item vectors given in Figure 11.

Insert Figure 11 about here

In Figure 11, one item, Item 4, is fairly easy and is measuring predominantly θ_1. That item has parameters $\mathbf{a}_4$ = (1.743, .153) and

$\underline{d}_4$ = 4.375. The second item, Item 5, is fairly difficulty and is measuring predominantly θ_2. The parameters for Item 5 are $\mathbf{a}_5$ = (.20, 1.638) and $\underline{d}_5$ = -3.960. The contour plots for the IRSs for these two items are given in Figures 12 and 13.

Insert Figures 12 and 13 about here

The question of interest here is "What will the SD be of the matrix of item scores generated by the interaction of a population of individuals with this two-item test?" To answer that question, a population of individuals must first be specified because the characteristics of the matrix are dependent on both the characteristics of the test and the characteristics of the examinee population. For this example, assume that the examinee population is distributed as a standard bivariate normal with ρ = 0.0. It is important for this example that the distribution be centered at (0, 0) in the

θ-space. If the examinee population has different location in the space, the results can be substantially different.

The equation for the .8-contour for Item 4 is

$$\theta_2 = -11.392\theta_1 - 19.536.$$

All θ_1,θ_2-points on this line will yield a probability of correct response of .8 when inserted into the equation for the item response model. In Case 1, the probability of correct response was the same on the second item when persons were on the equiprobable contour for the first item. To determine how much variation in probability of correct response for Item 5 occurs for persons on the .8-equiprobable contour for Item 4, the probability of correct response for Item 5 was computed for two points on the Item 4, .8-contour. When $\theta_1 = -1.5$, θ_2 is -2.448. For this point the probability of correct response to Item 5 is .0002. When $\theta_1 = -1.7$, θ_2 is -.17. For this point the probability of correct response to Item 5 is .0102. Thus, for these two points on the .8-contour

for Item 4, the probability of correct response to Item 5 is virtually the same.

It can easily be seen from inspection of the contour plots in Figures 12 and 13 that for all points on the equiprobable contours for Item 4 below a value of 1.0 on θ_2 there is little variation in the probability of correct response for Item 5. The same is true for Item 4 in the region of the space where θ_1 is greater than -1. Thus, for the majority of the population of examinees in this space, when the probability of correct response is constant on one item, it has very low variation on the other item, and vice versa. This implies that the same kind of mapping can be performed for Case 3 as was performed for Case 1. Persons who are located on the same equiprobable contour can be mapped onto a line that intersects the contours for both Item 4 and 5. A curved line that passes through (0, 0) and is orthogonal to the equiprobable contours will probably work well.

Figure 14 shows the estimated true score contours for the test composed of Items 4 and 5. The same line that is used to model the item

response data will also model the estimated true score contours with little loss of precision. Thus the data matrix generated by the interaction of the examinee population and these two items will have SD of one, even though the items emphasize quite different psychological dimensions.

Insert Figure 14 about here

This result is specific to the selection of a population centered at (0, 0) in the space. If a population of examinees that was centered at (-3, 3) was administered this same two-item test, the matrix of responses would have a SD of more than one because there would be substantial variation in correct response to Item 5 for persons on a equiprobable contour of Item 4.

Discussion

The purpose of this paper has been to discuss the conditions under which tests can be said to be unidimensional. The motivation for the paper is in part because unidimensionality assumptions are made quite often in the application of item response theory procedures. However, concerns over the dimensionality of test data is much broader than IRT applications. Many psychometric procedures implicitly assume unidimensionality even when no explicit statement of the assumption is made.

The first part of the paper deals with the distinction between psychological constructs and statistical variables. Typically when the dimensionality of test data is discussed, the statistical dimensionality is what is of concern. The discussions deal with whether a particular set of test data meet the unidimensionality assumptions of a particular psychometric procedure. It is critical that these discussions do not confuse these statistical considerations with the psychological interpretations of what the test is measuring. These are validity issues that are in many ways independent of the dimensionality issues.

The first part of the paper tries to make explicit that the dimensionality of test data really refers to the number of dimensions needed to summarize a data matrix that is the result of the interactions of a set of test items and a group of examinees. Only under very special conditions can the dimensionality assessed from the data matrix be said to apply to characteristics of either the set of test items or the group of examinees. Just because the data matrix can be modeled using one person variable does not mean that the people vary on only one dimension or that the test is only sensitive to differences on one dimension.

The second part of the paper summarizes the circumstances under which a set of test items will generate unidimensional data when responded to by a population of individuals. Two cases are presented. In the first case, all items are sensitive to the same combination of skills in the same way. This is shown by the parallelism of the equiprobable contour lines for all of the items in the test. That such tests yield unidimensional data when administered to a group of examinees was demonstrated empirically

in Reckase, Ackerman, and Carlson (1988). The important point here is that the test items need not be sensitive to only one psychological dimension for the interaction with the test to yield truly unidimensional data in the statistical sense.

The second situation in which unidimensional data are generated by the interaction of a set of test items with a set of people, is when the psychological dimensions are strongly confounded with the difficulty of the test items. This case yields unidimensional data because there is little variation in the probability of correct response on the items measuring other dimensions when there is little variation in probability of correct response for items measuring the first dimension. This result has been demonstrated empirically by Davey, Ackerman, Reckase & Spray (1989).

While these two cases both yield statistically unidimensional data, the first results in number correct scores that have the same substantive meaning throughout the score scale and the second yields a number correct score scale that shifts in its meaning from low performance to high

performance. The first case is an example of obtaining truly unidimensional data from a test measuring multiple psychological constructs, while the second is a case of obtaining data that appears unidimensional from a statistical perspective but which has a multidimensional psychological meaning.

These results both simplify and complicate the analysis of test data. On the one hand, the application of psychometric procedures is simplified because the results imply that tests do not have to measure narrowly defined, pure psychological traits for procedures that assume unidimensionality to apply. They need only measure the same combination of traits. On the other hand, the fact that the meaning of a score scale can change with the level of performance greatly complicates the interpretation of the psychological constructs underlying a test. In both cases, two tests can be considered truly parallel only when the unidimensional score scales have the same orientation in the multidimensional psychological trait space.

References

Davey, T., Ackerman, T. A., Reckase, M. D., & Spay, J. A. (1989, July). Interpreting score differences when item difficulty and discrimination are confounded. Paper presented at the meeting of the Psychometric Society, Los Angeles.

Embretson, S. E. (1985). Multicomponent models for test design. In S. E. Embretson (Ed.) Test design: Developments in psychology and psychometrics. New York: Academic Press.

Goldstein, H. & Wood, R. (1989). Five decades of item response modelling. British Journal of Mathematical and Statistical Psychology, **42**, 139-167.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.) Fducational Measurement (3rd Ed.). New York: American Council on Education and Macmillan.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). Item response theory: Application to psychological measurement. Homewood, IL: Dow Jones-Irwin.

Linn, R. L. (1989). Educational Measurement (3rd Ed.). New York: American Council on Education and Macmillan.

Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Reckase, M. D. (1989, September). The interpretation and application of multidimensional item response theory models; and computerized testing in the instructional environment: Final Report (Research Report ONR 89-2). Iowa City, IA: The American College Testing Program.

Reckase, M. D., Ackerman, T. A. & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. Journal of Educational Measurement, **25(3)**, 193-204.

Reckase, M. D. & McKinley, R. L. (1985). Some latent trait theory in a multidimensional latent space. In D. J. Weiss (Ed.) Proceedings of the 1982 item response theory and computerized testing conference. Minneapolis: University of Minnesota.

Figure Captions

Figure 1. Perspective plot of the item response surface for Item 1.

Figure 2. Contour plot of the item response surface for Item 1.

Figure 3. Vector plot for Item 1.

Figure 4. Perspective plot of the item response surface for Item 2.

Figure 5. Contour plot of the item response surface for Item 2.

Figure 6. Contour plot of the estimated true score surface for a test composed of Items 1 and 2.

Figure 7. Vector plot for Items 1 and 2.

Figure 8. Contour plot of the item response surface for Item 3.

Figure 9. Contour plot of the estimated true score surface for a test composed of Items 1 and 3.

Figure 10. Vector plot for Items 1 and 3.

Figure 11. Vector plot for Items 4 and 5.

Figure 12. Contour plot of the item response surface for Item 4.

Figure 13. Contour plot of the item response surface for Item 5.

Figure 14. Contour plot of the estimated true score surface for a test composed of Items 4 and 5.

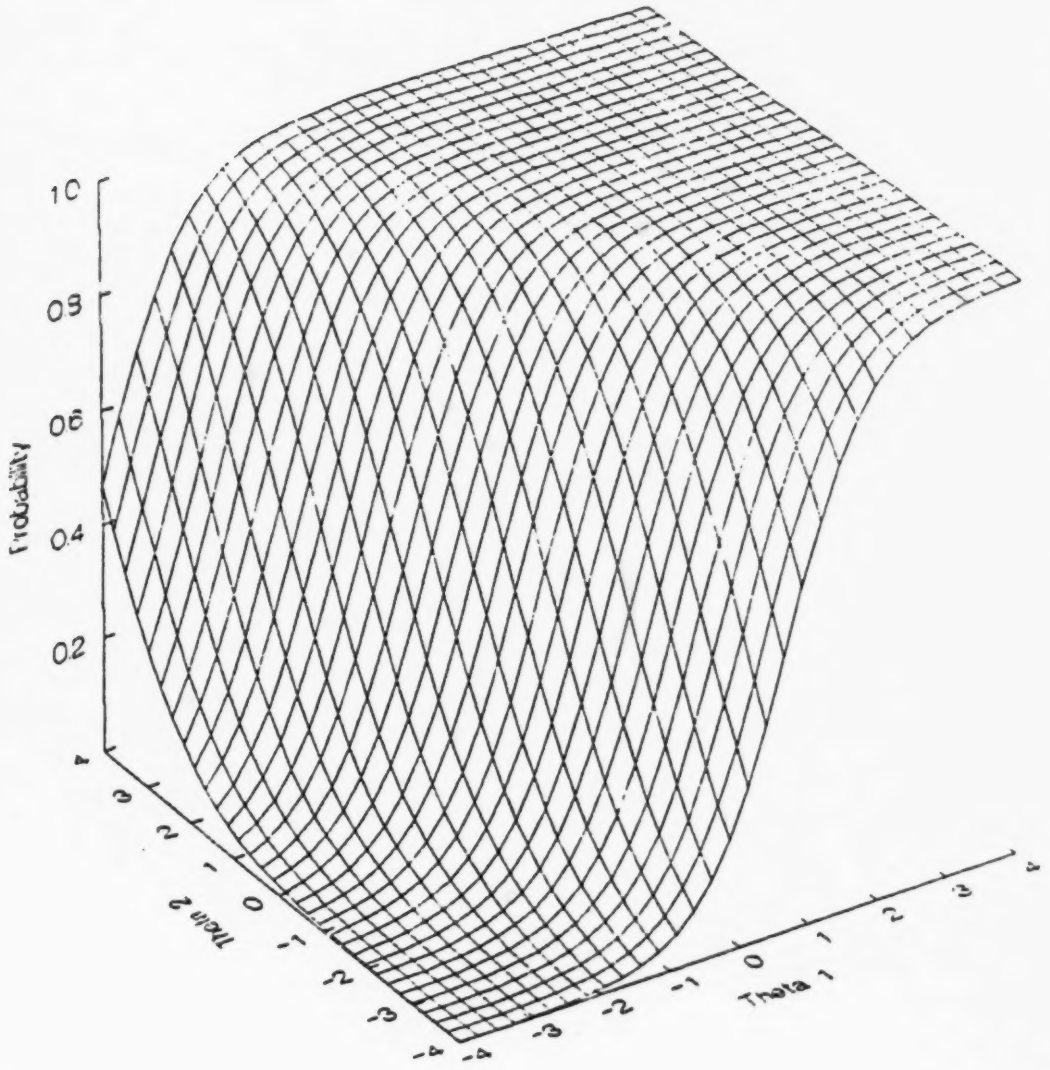

Figure 1: Perspective plot of the item response surface for Item 1.

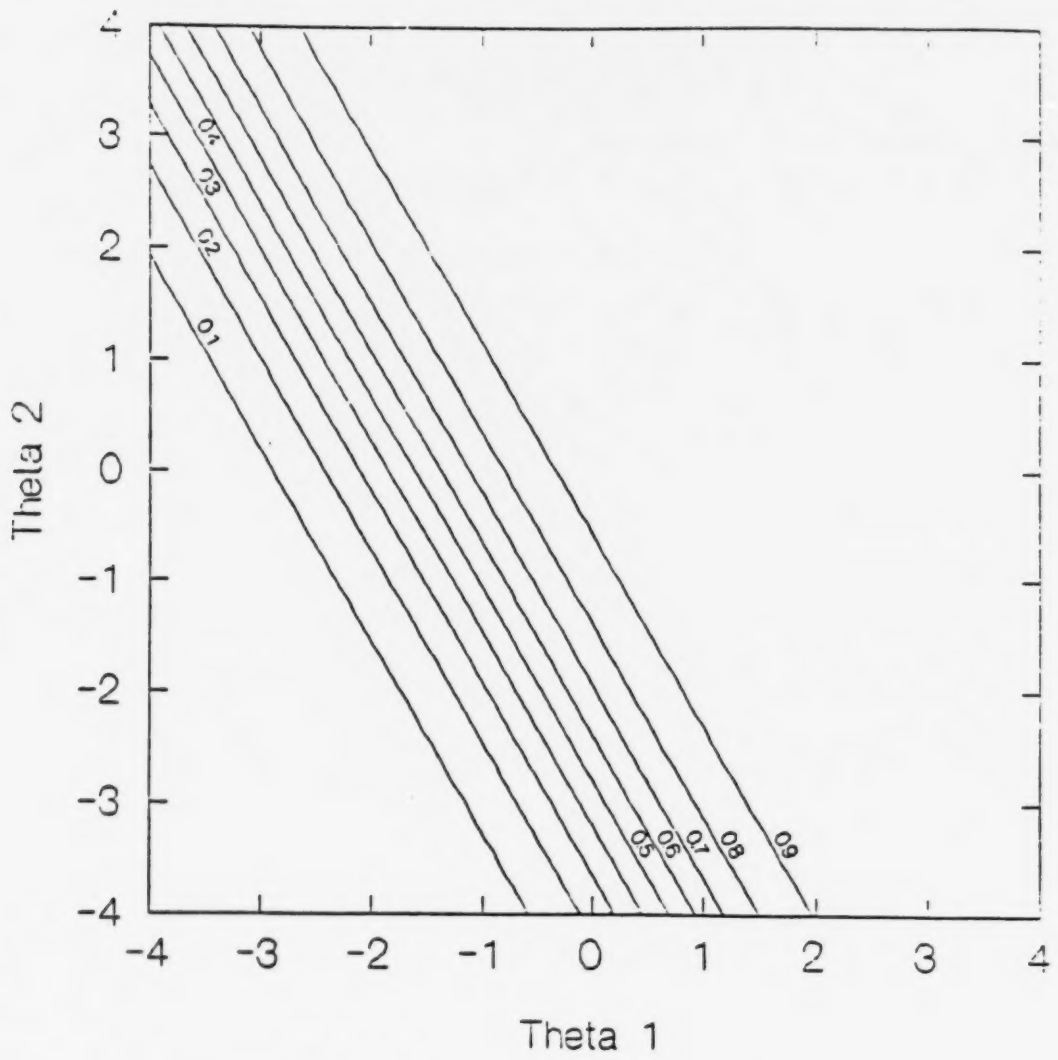

Figure 2: Contour plot of the item response surface for Item 1.

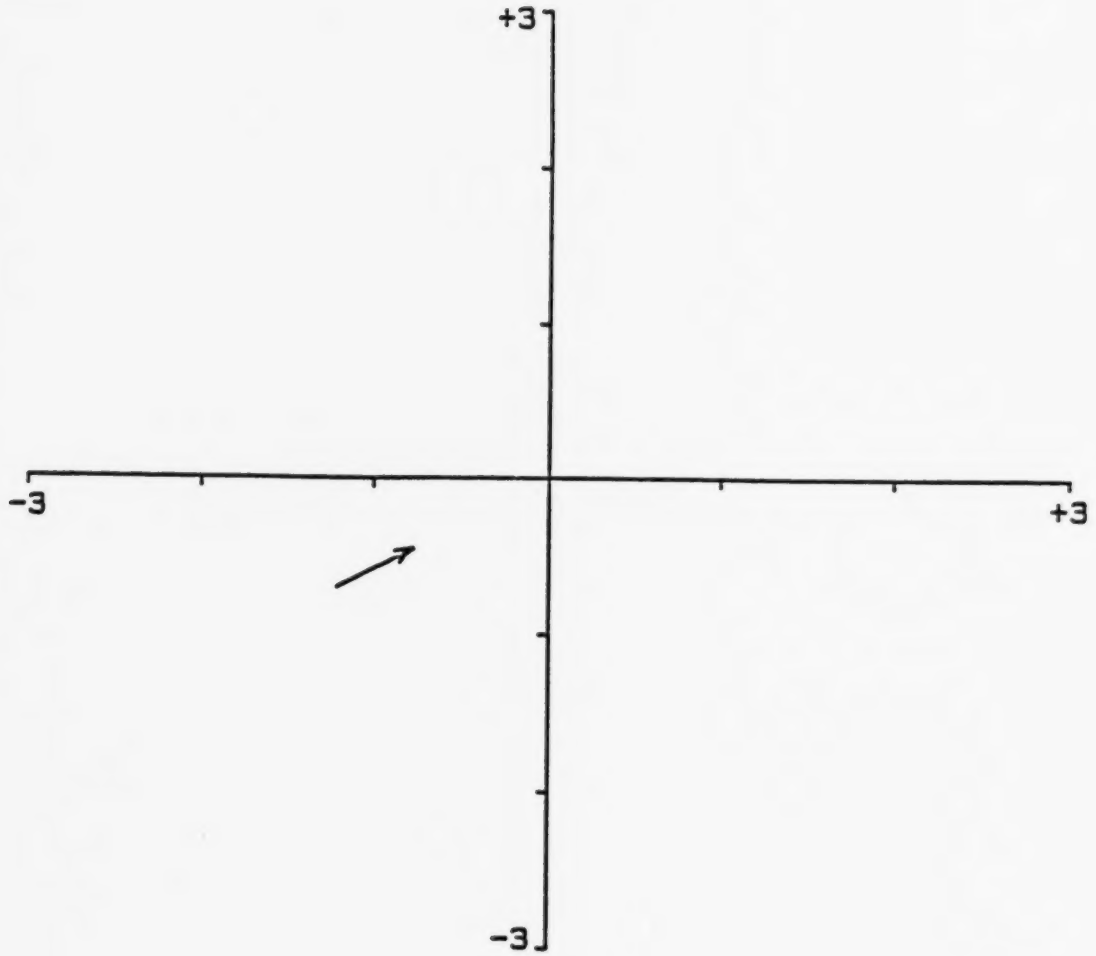

Figure 3: Vector plot for Item 1.

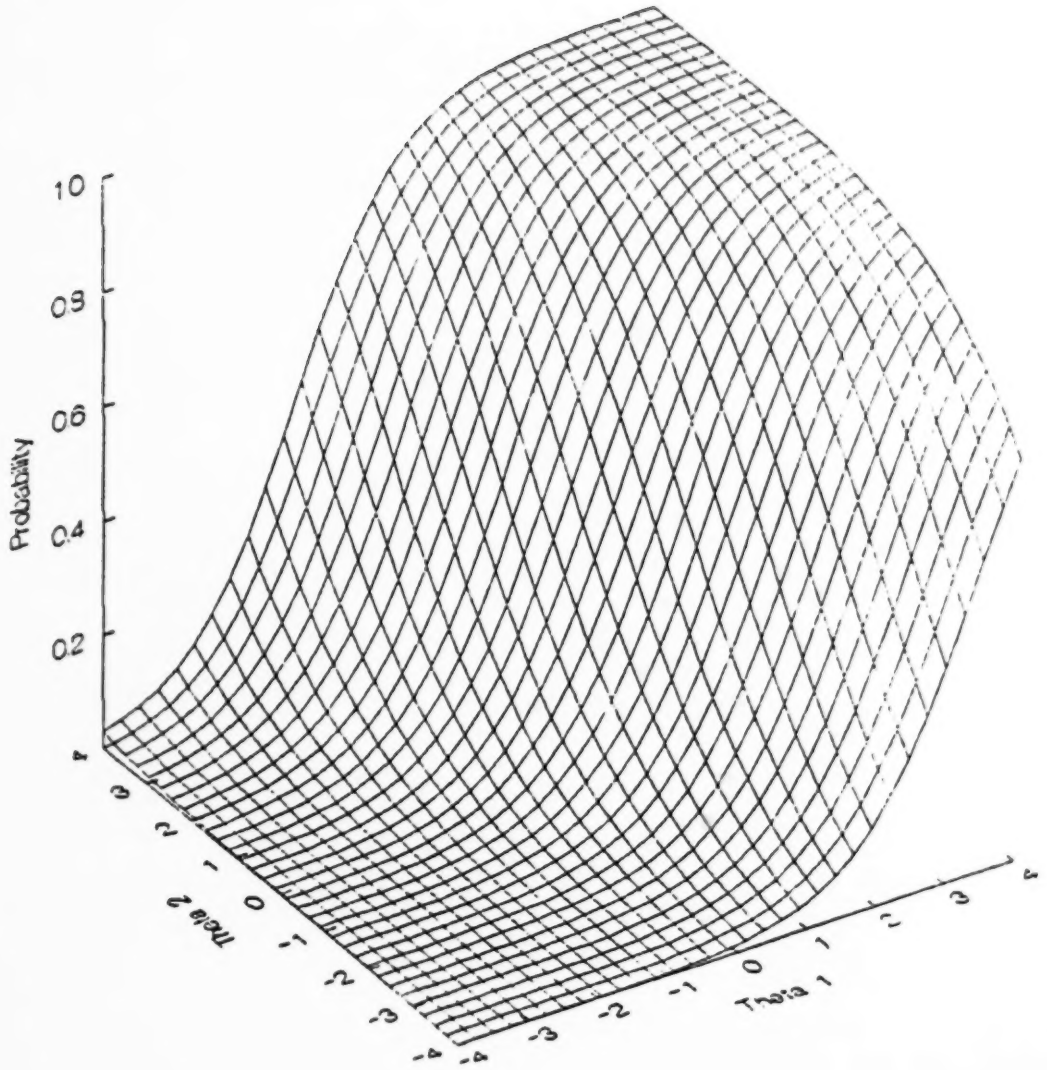

Figure 4: Perspective plot of the item response surface for Item 2.

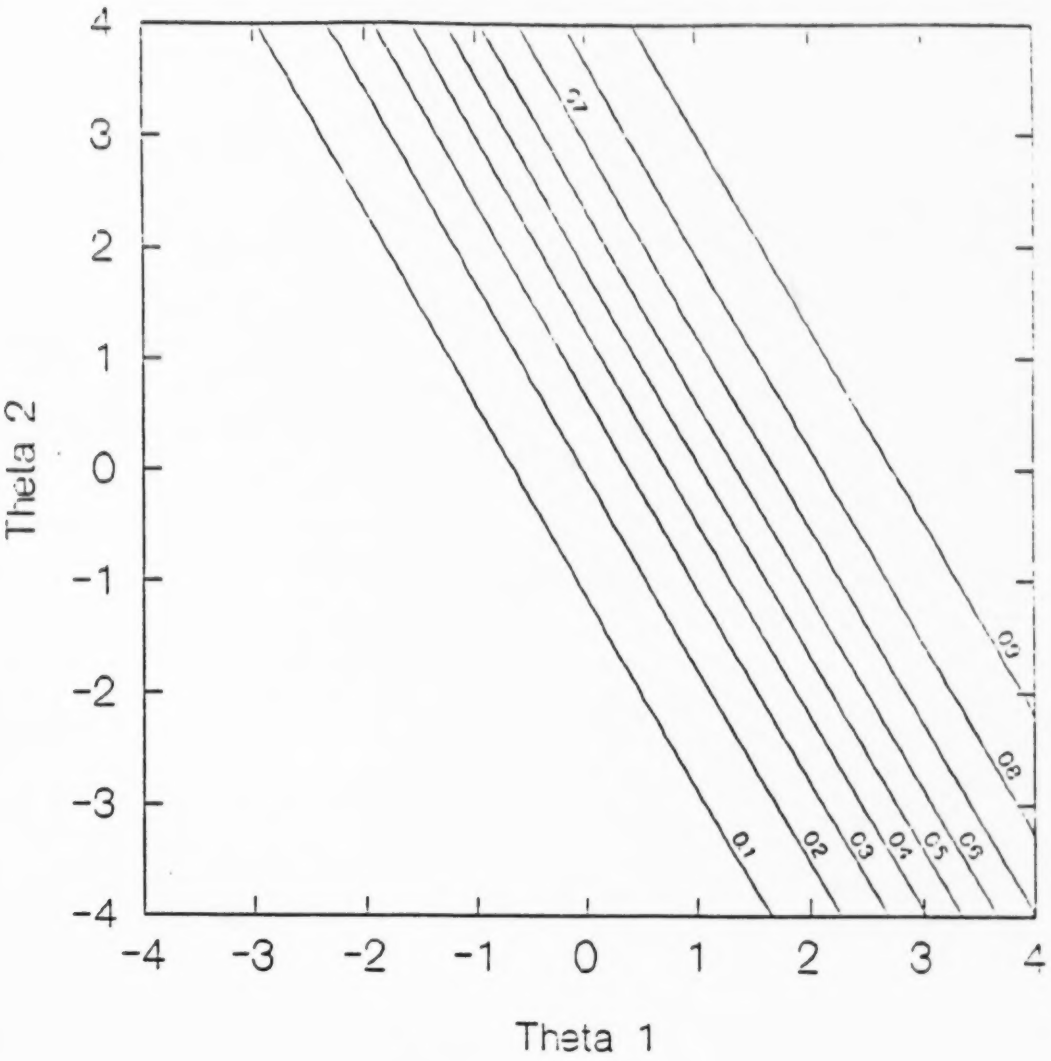

Figure 5: Contour plot of the item response surface for Item 2.

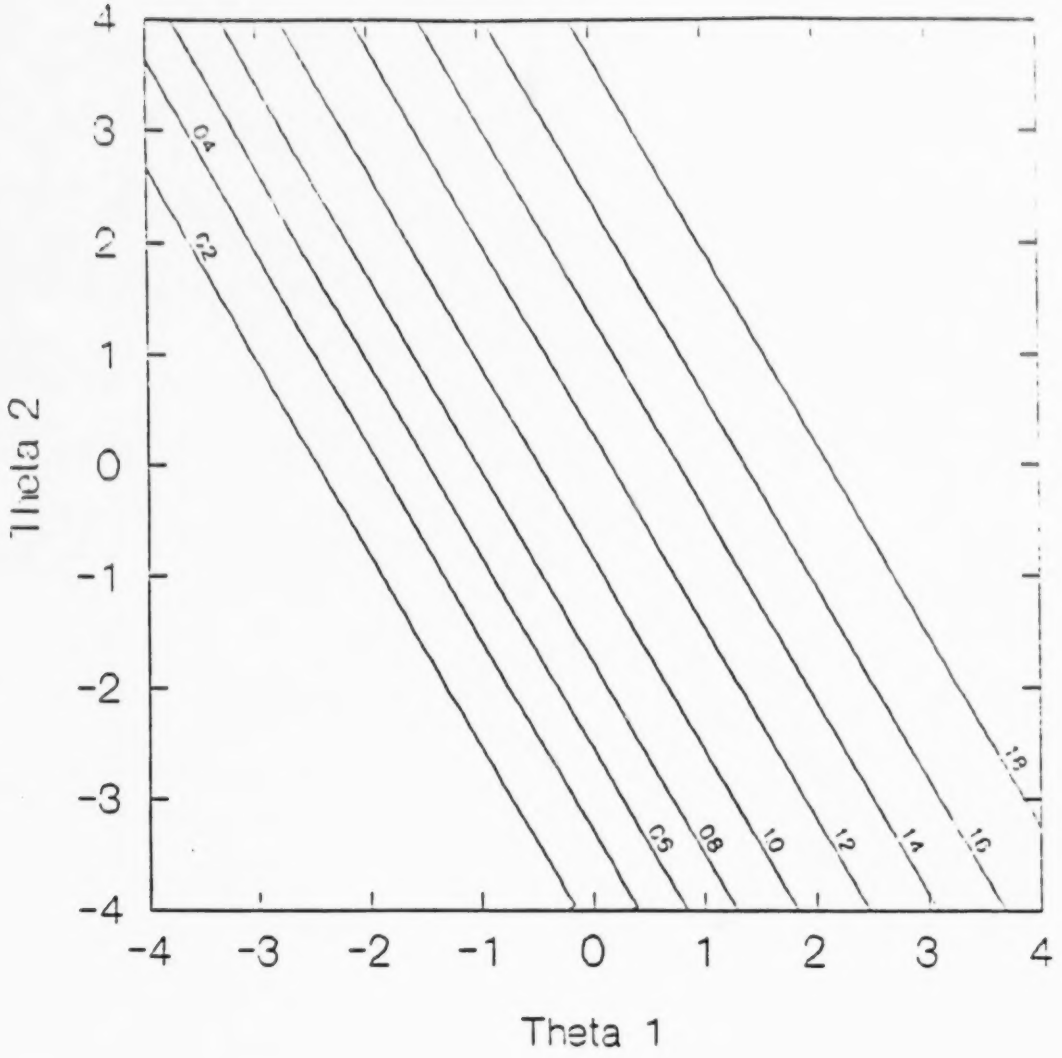

Figure 6: Contour plot of the estimated true score surface for a test composed of Items 1 and 2.

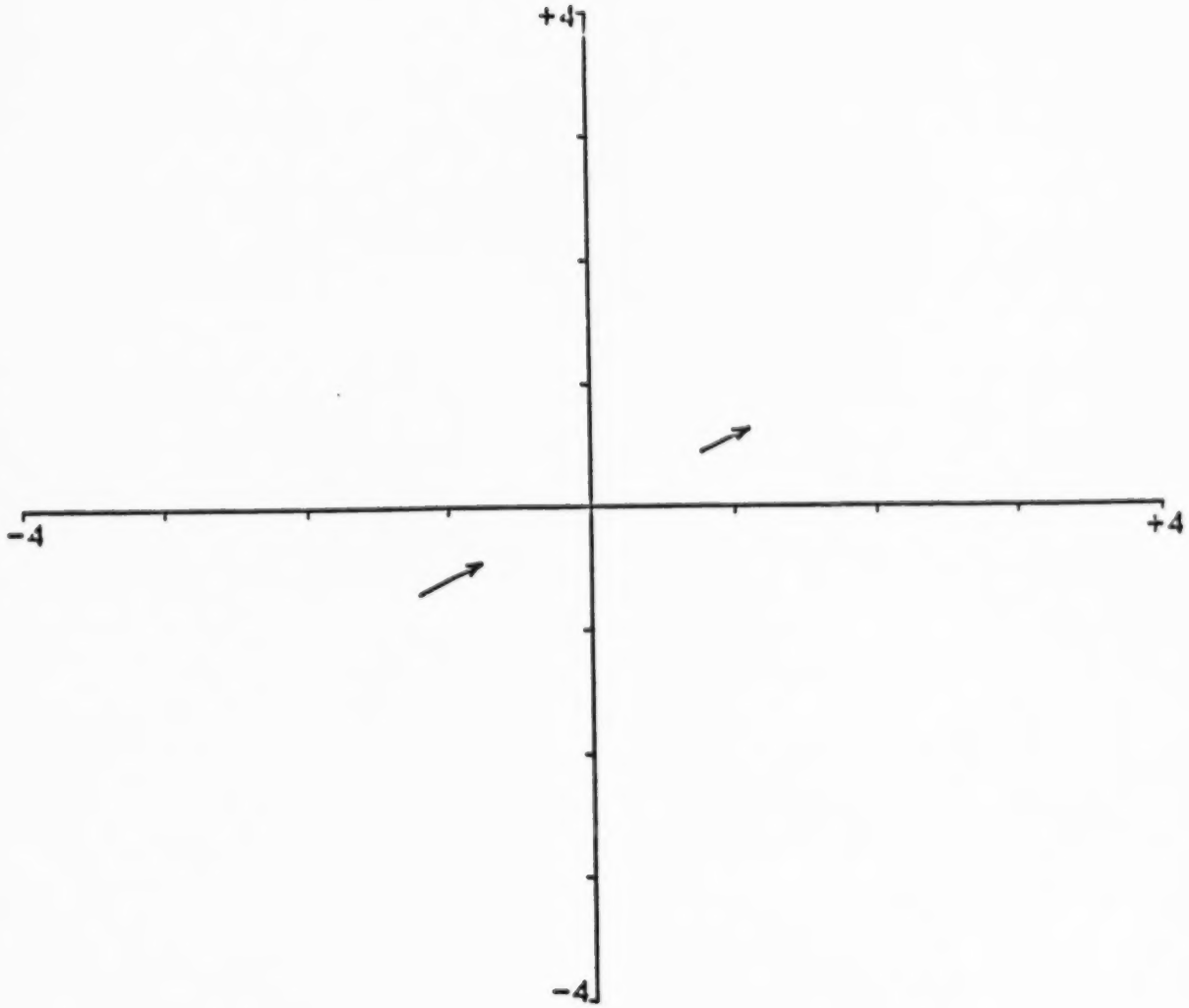

Figure 7: Vector plot for Items 1 and 2.

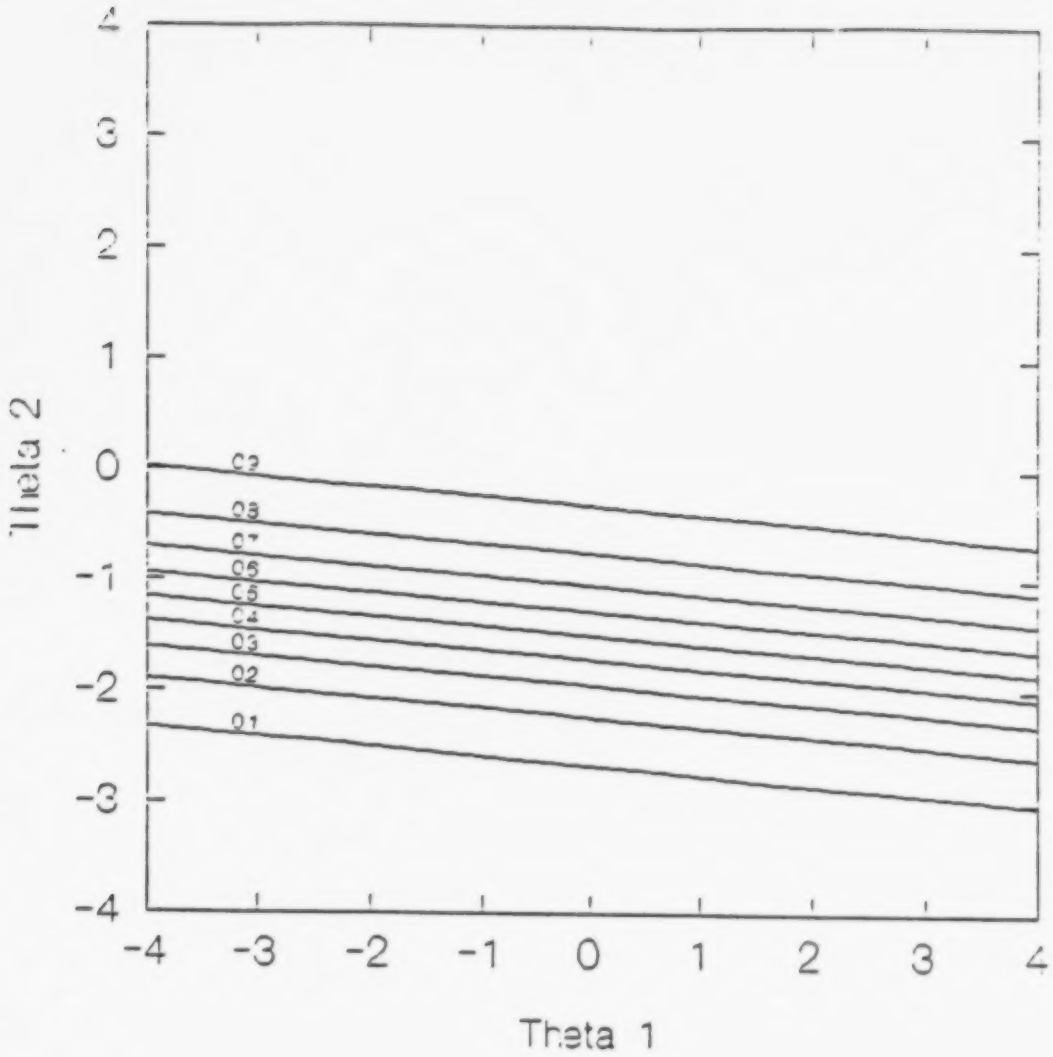

Figure 8: Contour plot of the item response surface for Item 3.

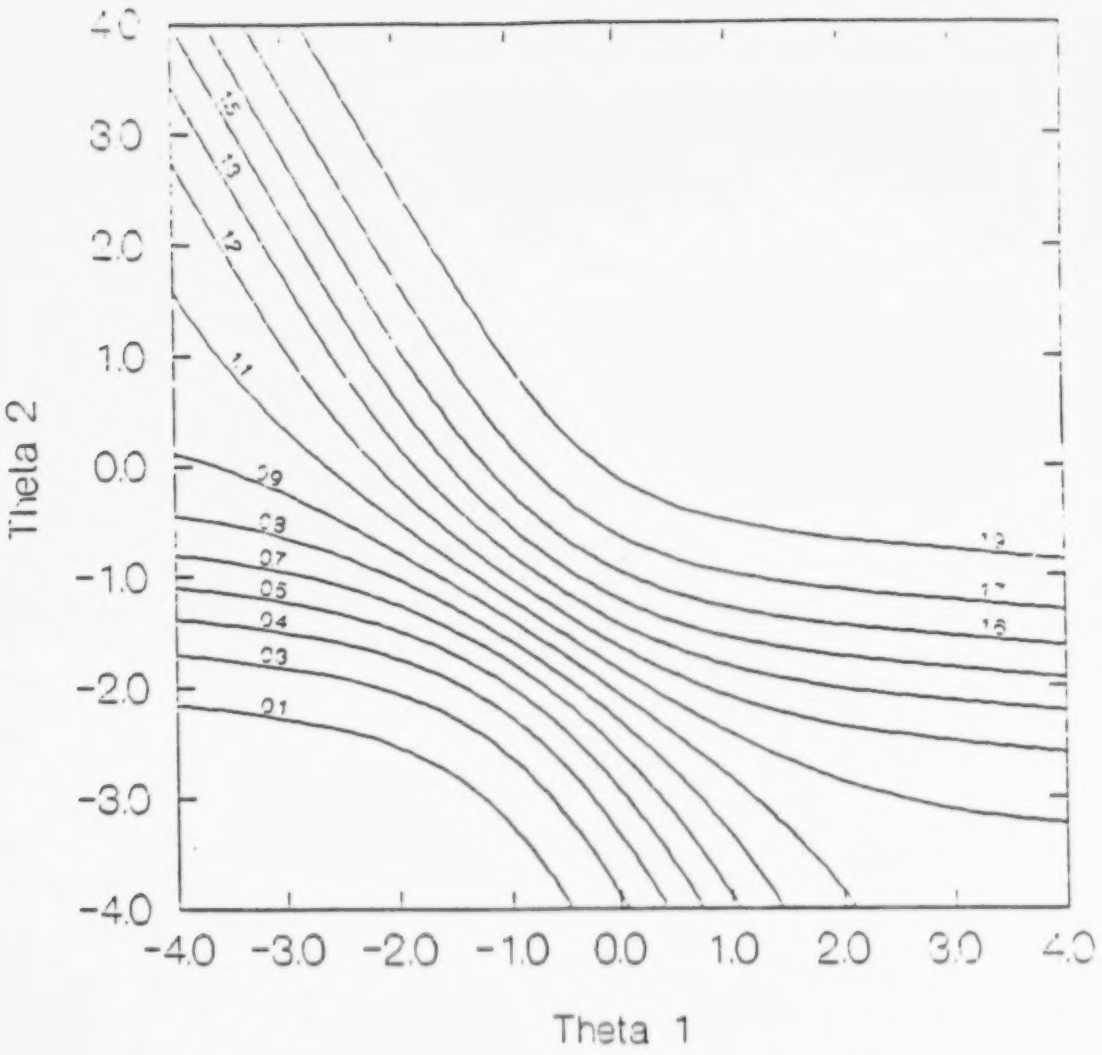

Figure 9: Contour plot of the estimated true score surface for a test composed of Items 1 and 3.

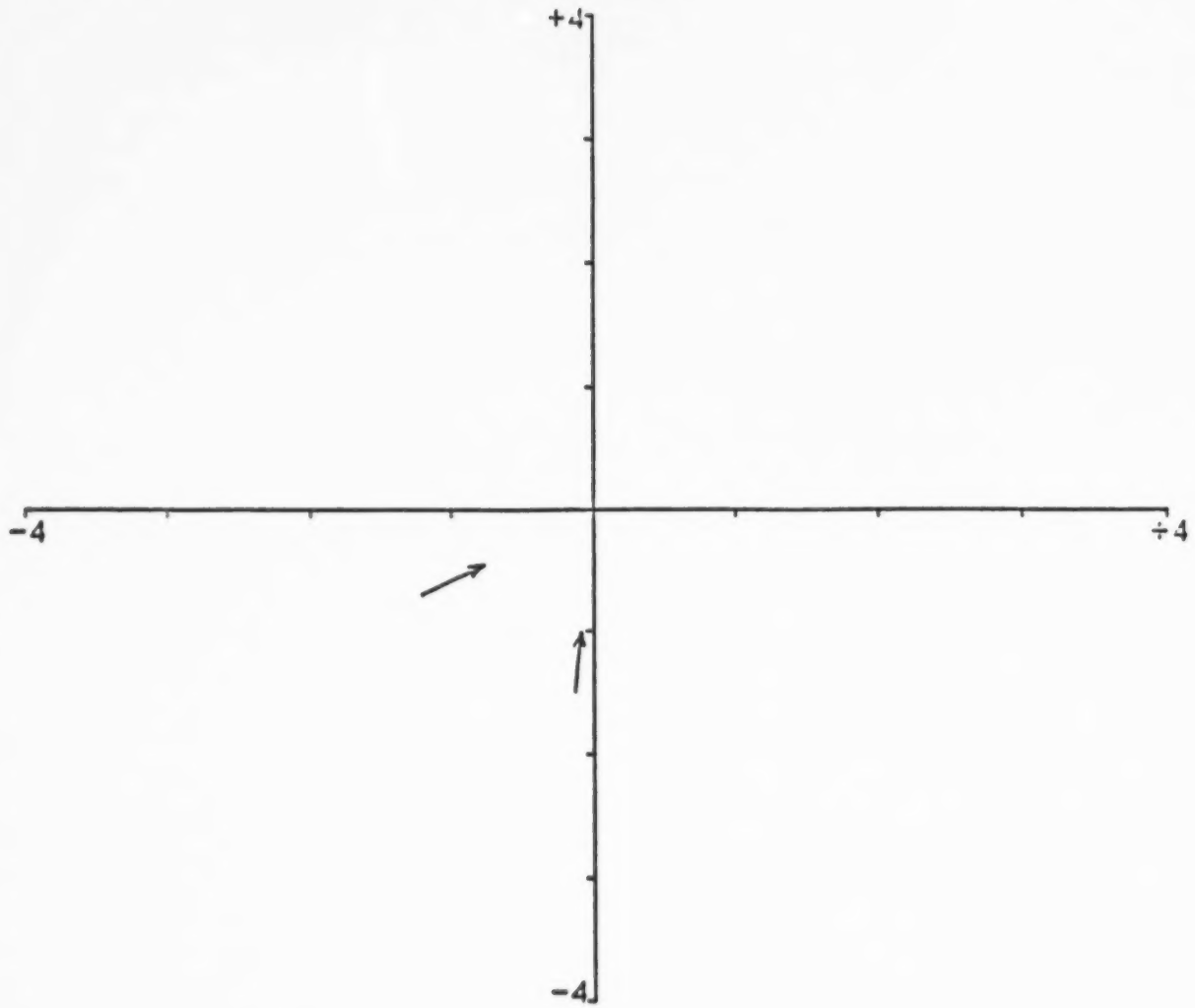

Figure 10: Vector plot for Items 1 and 3.

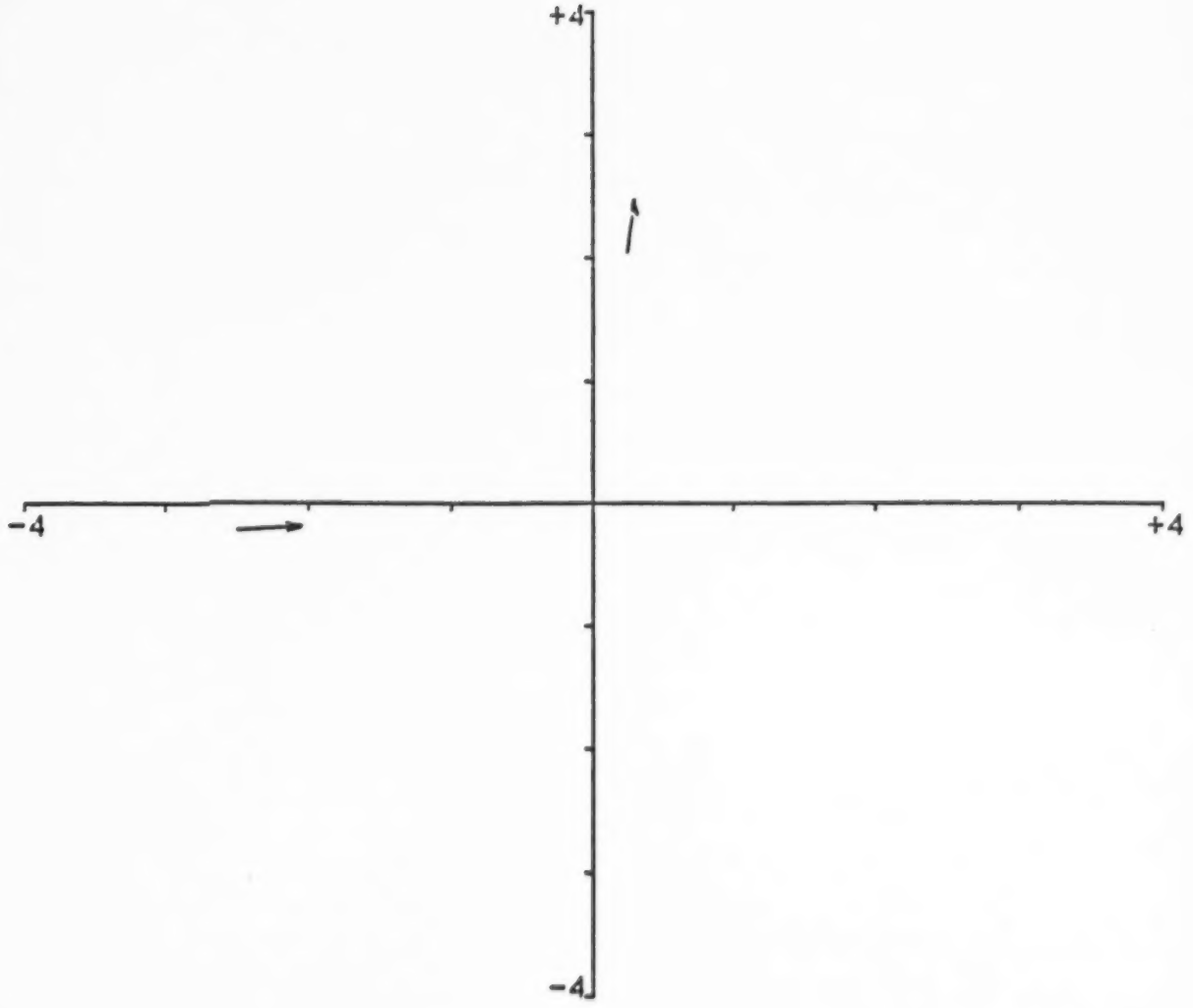

Figure 11: Vector plot for Items 4 and 5.

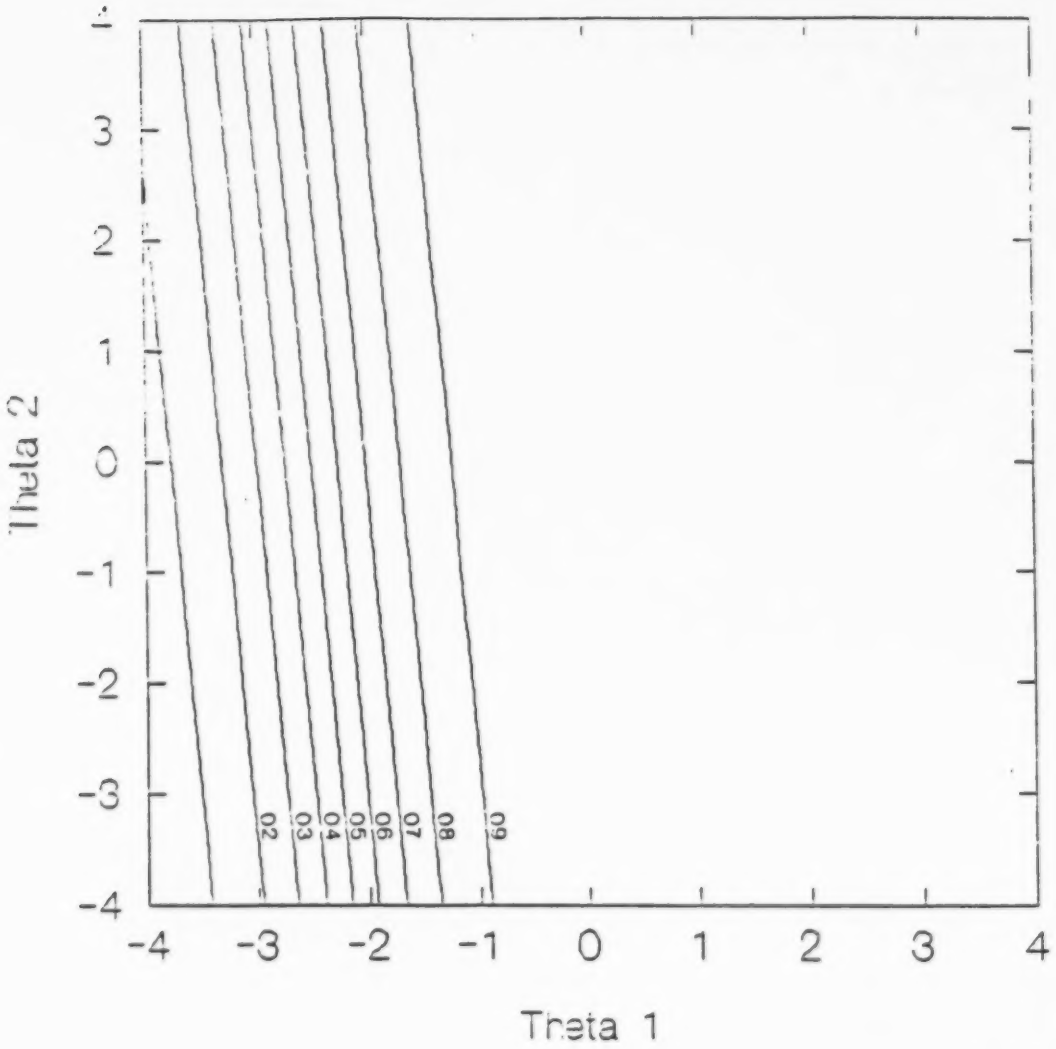

Figure 12: Contour plot of the item response surface for Item 4.

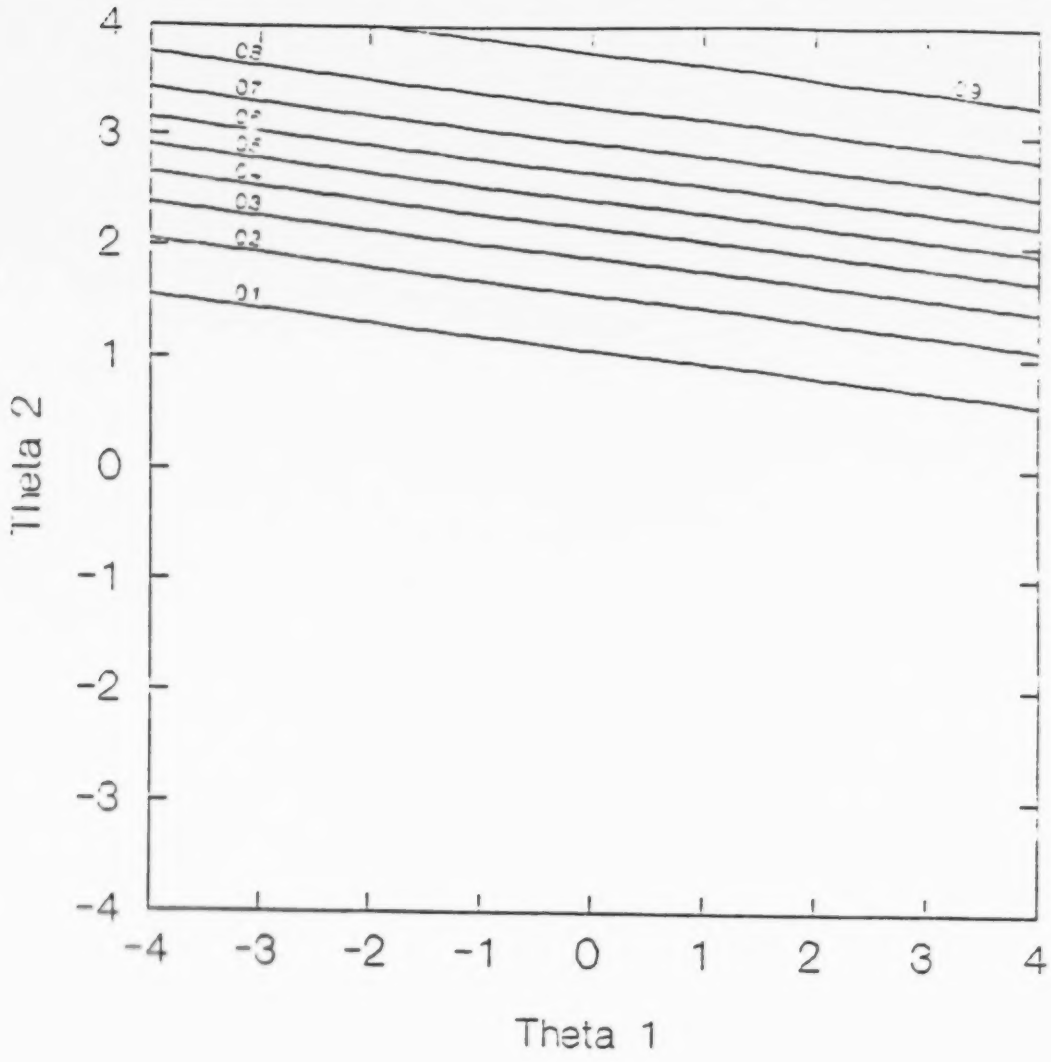

Figure 13: Contour plot of the item response surface for Item 5.

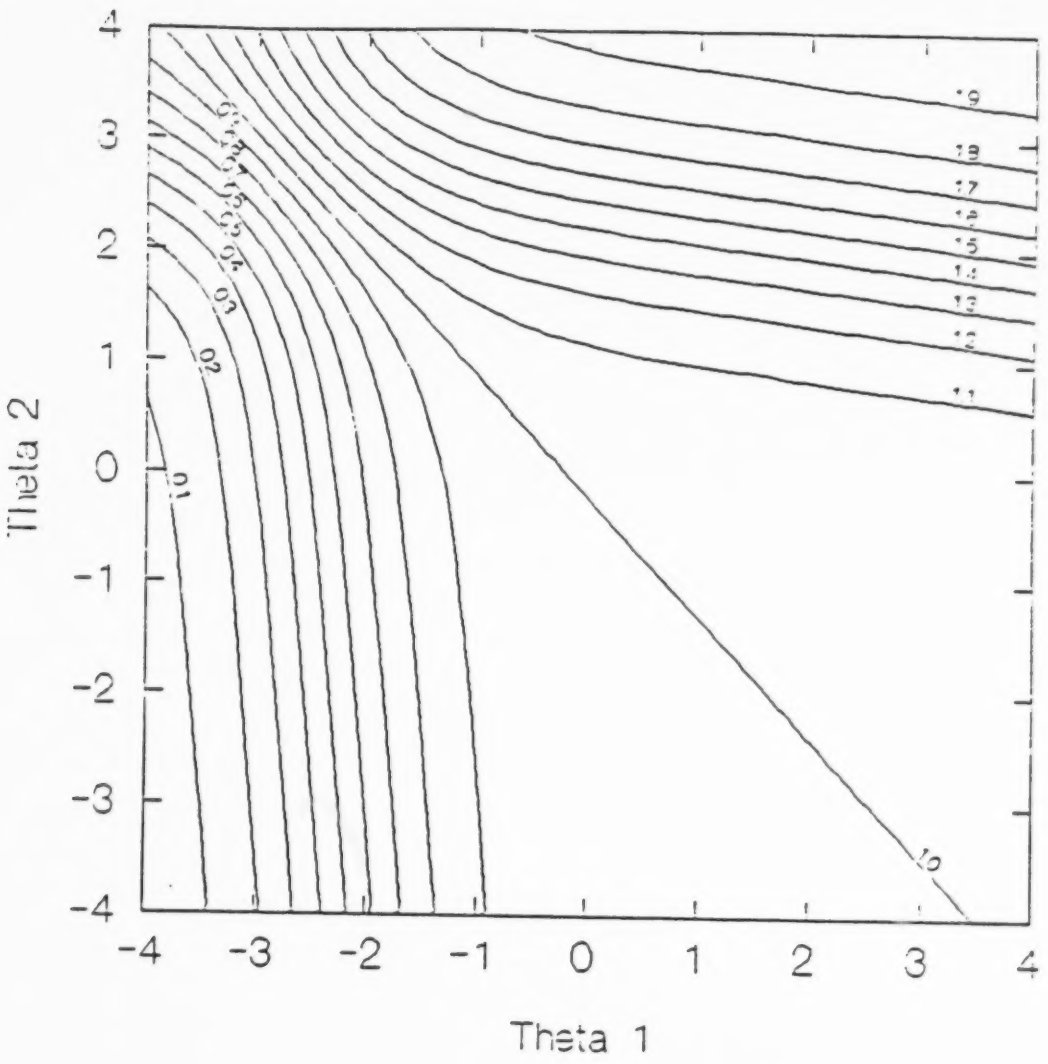

Figure 14: Contour plot of the estimated true score surface for a test composed of Items 4 and 5.